Econ 4130 Exam 2009 H - POSTPONED

Sketch answers in << ….. >>.

Problem 1

Introduction. Suppose that the joint distribution of X_1, X_2, X_3 is trinomial (i.e., multinomial with three categories - *see the appendix at the end of the problem set for a summary of the multinomial distribution*), with parameters *n* (number of trials) and probabilities, p_1, p_2, p_3, where $X_1 + X_2 + X_3 = n$, and $p_1 + p_2 + p_3 = 1$.

Here X_j is the number of occurrences of category *j* in *n* independent trials where, in each trial, the probability of category *j* occurring is p_j for $j = 1, 2, 3$.

Without further restrictions on the parameters, the maximum likelihood estimators (mle) of p_1, p_2, p_3 are given by $\hat{p}_j = \frac{X_j}{n}$, for $j = 1, 2, 3$ (you do not need to show this here).

Since X_3 is determined by $X_3 = n - X_1 - X_2$, we may also consider the trinomial distribution as a two-dimensional joint distribution of the random vector, $(X_1, X_2)'$. Using the information in the appendix, the covariance matrix of $(X_1, X_2)'$ is given by

$$\operatorname{cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 \\ -np_1p_2 & np_2(1-p_{2)} \end{pmatrix}$$

A. Explain why the covariance matrix of $(\hat{p}_1, \hat{p}_2)'$ can be written

(1) $$\operatorname{cov}\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} = \frac{1}{n} C, \quad \text{where} \quad C = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) \end{pmatrix}$$

<< Answer:

$$\operatorname{cov}\left(\hat{p}_i, \hat{p}_j\right) = E\left[\left(\frac{X_i}{n} - E\frac{X_i}{n}\right)\left(\frac{X_j}{n} - E\frac{X_j}{n}\right)\right] = E\left[\left(\frac{X_i}{n} - \frac{EX_i}{n}\right)\left(\frac{X_j}{n} - \frac{EX_j}{n}\right)\right] =$$

$$= \frac{1}{n^2}\operatorname{cov}\left(X_i, X_j\right)$$

Since a common scalar factor inside a matrix can be taken as a single scalar factor outside the matrix, the result follows. >>

B. **(i)** Without referring to the general properties of mle's, show that $\hat{p}_j$ is a consistent estimator of p_j. [**Hint:** Use, e.g., Chebyshev's inequality.]

(ii) Find a consistent estimator of the matrix *C* in (1), and explain why it is consistent. [**Hint:** Remember (you do not need to prove it here) that a matrix of estimators is consistent if and only if each element in the matrix is consistent]

<<**Answer:** **(i)** For $\varepsilon > 0$ arbitrary, we have since $E(\hat{p}_j) = p_j$,

$$P(|\hat{p}_j - p_j| > \varepsilon) \le \frac{\operatorname{var}(\hat{p}_j)}{\varepsilon^2} = \frac{p_j(1-p_j)}{n\varepsilon^2} \underset{n\to\infty}{\to} 0$$

(ii) All elements in *C* are continuous functions of p_1, p_2. Using the continuity property of consistency, substituting p_j by $\hat{p}_j$ in all elements of *C* will produce consistent estimators of the elements. >>

C. Leaving the trinomial distribution for a moment, let *X* and *Y* be two arbitrary random variables (with finite expected values and variances) such that the regression function of *Y* with respect to *X* is a linear function of *x*, i.e.,

$$E(Y \mid x) = \alpha + \beta x \quad \text{(where } \alpha, \beta \text{ are constants)}$$

Show that the assumption of linearity of the regression function implies that the regression coefficient, β, must be given by

$$\beta = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$$

[**Hint:** Utilize the law of total expectation.]

<< **Answer:** Note that the proof of this is given in the solution to the no-seminar week exercise for week 39 on the net – so the main challenge for the candidate is to remember this. Otherwise:

Write $E(X)=\mu_X$, $\text{var}(X)=\sigma_X^2$, $E(Y)=\mu_Y$. Then

$\mu_Y = E\big[E(Y\mid X)\big] = E\big[\alpha+\beta X\big] = \alpha+\beta\mu_X$ and

$E\big(XY\big) = E\big[X\cdot E(Y\mid X)\big] = E\big[\alpha X+\beta X^2\big] = \alpha\mu_X+\beta(\sigma_X^2+\mu_X^2)$

Hence

$\text{cov}(X,Y) = E(XY)-E(X)E(Y) = \alpha\mu_X+\beta(\sigma_X^2+\mu_X^2)-\mu_X(\alpha+\beta\mu_X) =$

$= \beta\sigma_X^2$

which gives $\beta = \dfrac{\text{cov}(X,Y)}{\text{var}(X)}$ >>

D. We now return to the trinomial distribution in section **A:**

(i) Show that the conditional distribution of X_1 given that $X_2=x_2$, is binomial with parameters, $n-x_2$ (trials) and (success probability), $\dfrac{p_1}{1-p_2}$. [**Hint:** Use the definition of the conditional pmf (probability mass function).]

(ii) Set up an expression for the regression function of X_1 with respect to X_2, i.e., $E(X_1\mid x_2)$.

<< **Answer: (i) :**

$$P(X_1=x_1\mid X_2=x_2)=\frac{P(X_1=x_1\cap X_2=x_2)}{P(X_2=x_2)}=\frac{\dfrac{n!}{x_1!x_2!(n-x_1-x_2)!}p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2}}{\dfrac{n!}{x_2!(n-x_2)!}p_2^{x_2}(1-p_2)^{n-x_2}}=$$

$$=\frac{(n-x_2)!}{x_1!(n-x_2-x_1)!}\left(\frac{p_1}{1-p_2}\right)^{x_1}\left(1-\frac{p_1}{1-p_2}\right)^{n-x_2-x_1}$$

which is the pmf of the $\text{Bin}\,(n-x_2,\dfrac{p_1}{1-p_2})$ - distribution.

(ii) Hence $E\big(X_1\mid x_2\big)=(n-x_2)\dfrac{p_1}{1-p_2}$ (which is linear in x_2). >>

E. Show that $\text{cov}(X_1, X_2) = -np_1p_2$. [**Hint:** Use section **C** and **D**.]

<< **Answer:** From **Dii** the regression coefficient in the linear $E(X_1 \mid x_2)$ is

$\beta = -\dfrac{p_1}{1-p_2}$, and **C** gives:

$\text{cov}(X_1, X_2) = \beta\sigma^2_{X_2} = -\dfrac{p_1}{1-p_2} np_2(1-p_2) = -np_1p_2$ >>

F. Now, assume that the parameters of our trinomial model satisfy the following restriction: $p_1 = p_2$. Call the common (unknown) value, θ, so that

$$p_1 = \theta, \quad p_2 = \theta, \quad p_3 = 1-2\theta$$

In addition we assume that $0 < \theta < 1/2$.

Show that the mle for θ is: $\hat{\theta} = \dfrac{X_1 + X_2}{2n}$

<< **Answer:** Let *c* denote a constant that does not depend on θ. The log likelihood becomes

$$\ell(\theta) = c + x_1 \ln(\theta) + x_2 \ln(\theta) + (n - x_1 - x_2)\ln(1-2\theta) = c + (x_1 + x_2)\ln(\theta) + (n - x_1 - x_2)\ln(1-2\theta)$$

and

$$\ell'(\theta) = \frac{x_1 + x_2}{\theta} - 2\frac{n - x_1 - x_2}{1-2\theta} = \frac{(x_1 + x_2) - 2\theta(x_1 + x_2) - 2\theta(n - x_1 - x_2)}{\theta(1-2\theta)} = \frac{x_1 + x_2 - 2n\theta}{\theta(1-2\theta)}$$

which is 0 for $\theta = \hat{\theta} = \dfrac{x_1 + x_2}{2n}$. Clearly $\ell'(\theta) > 0$ for $\theta < \hat{\theta}$ and $\ell'(\theta) < 0$ for $\theta > \hat{\theta}$ showing that we have a maximum. >>

G. It can be shown (you do not need to do that here) that $Y = X_1 + X_2$ is binomially distributed with parameters $(n, 2\theta)$. In general, if *Y* is binomially distributed (n, q), then, according to the theory for binomial distributions,

$\sqrt{n}\left(\hat{q} - q\right) \underset{n\to\infty}{\overset{D}{\to}} Z \sim N\left(0,\ q(1-q)\right)$ where $\hat{q} = Y/n$. Utilize this to develop a formula for an approximate $1-\alpha$ confidence interval for θ for large n.

<< **Answer:** We have $Y = X_1 + X_2$ is binomial with parameters $(n, q) = (n, 2\theta)$. The usual argument based on Slutsky and the normal limit distribution now gives the approximate $1-\alpha$ CI for q, using the mle $\hat{q} = Y/n = 2\hat{\theta}$:

$$\hat{q} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$$

Hence

$$1-\alpha \approx P\left(\hat{q} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \le q \le \hat{q} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \right) =$$

$$= P\left(\hat{q} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \le 2\theta \le \hat{q} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \right) =$$

$$= P\left(\frac{\hat{q}}{2} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{4n}} \le \theta \le \frac{\hat{q}}{2} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{4n}} \right) =$$

$$= P\left(\hat{\theta} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{4n}} \le \theta \le \hat{\theta} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{q}(1-\hat{q})}{4n}} \right)$$

which gives the CI for θ >>

Problem 2

In December 2009, according to VG "nett-avis" and the "Sentio Research Group" opinion poll institute, 49.2% of a random sample of Norwegians were against applying for membership in the European Union (EU), 38.8% were for EU, and the remaining percentage, 12%, of the sample had no opinion. In this problem we are interested in 95% error margins for the difference between the percentage against EU and the percentage in favour of EU in the population. Unfortunately, VG did not report neither error margins nor sample size. For simplicity we will calculate a 95% confidence interval (CI) for the difference based on the assumption that the sample is a simple random sample of size,

$n = 1000$ (a sample size often used by opinion poll institutes). In that case we may take the trinomial distribution as a model for the number of occurrences of the three categories, "against EU", "in favour of EU", and "no opinion".

Thus, let X_1, X_2, X_3 denote the number of the categories, "against", "in favour", and "no opinion", respectively, occurring in the sample. Assuming $n = 1000$, the observed values are $X_{1,obs} = 492, \quad X_{2,obs} = 388, \quad X_{3,obs} = 120$. As our model we assume that X_1, X_2, X_3 is trinomially distributed with parameters $n = 1000$ and probabilities p_1, p_2, p_3, where p_j represents the relative frequency of category *j* in the population.

A. According to general asymptotic theory (which you do not need to show here), we can, for large *n* (which is the case here), base our inference on the approximate multivariate normal distribution for the mle, $(\hat{p}_1, \hat{p}_2)' = \left(X_1/n,\ X_2/n\right)'$, given by:

(2) $$\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix} \overset{approximately}{\sim} N\left(\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \frac{1}{n}\hat{C}_{obs} \right)$$

where $\hat{C}_{obs}$ is the observed value of a consistent estimate of *C* defined in problem 1 **A.**

(i) Calculate $\hat{C}_{obs}$. **(ii)** Calculate an approximate 95% CI for the difference $p_1 - p_2$.

<< **Answer:** **(i)** The observed values of the mle's are: $\hat{p}_1 = 0.492, \quad \hat{p}_2 = 0.388$,

giving: $\hat{C}_{obs} = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & -\hat{p}_1\hat{p}_2 \\ -\hat{p}_1\hat{p}_2 & \hat{p}_2(1-\hat{p}_2) \end{pmatrix}_{obs} = \begin{pmatrix} 0.249936 & -0.190896 \\ -0.190896 & 0.237456 \end{pmatrix} \approx 1000 \cdot \mathrm{cov}\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix}$

(ii) We get $\mathrm{var}(\hat{p}_1 - \hat{p}_2) = \mathrm{var}(\hat{p}_1) + \mathrm{var}(\hat{p}_2) - 2\mathrm{cov}(\hat{p}_1, \hat{p}_2) \approx 0.000869184$

The standard error is: $\mathrm{SE}(\hat{p}_1 - \hat{p}_2) \approx \sqrt{0.000869184} = 0.02948$ and the estimate is

$\hat{p}_1 - \hat{p}_2 = 0.104$. Hence the approximate 95% CI for $p_1 - p_2$ is

$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \mathrm{SE}(\hat{p}_1 - \hat{p}_2) = 0.104 \pm (1.96)(0.02984) = [0.046,\ 0.162]$

(**Alternatively** we could use the matrix version in lecture notes to chapter 8 on the net,

using $\hat{p}_1 - \hat{p}_2 = (1,-1)\begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \end{pmatrix}$, and calculate the variance by

$\mathrm{var}(\hat{p}_1 - \hat{p}_2) \approx \frac{1}{n}(1,-1)\hat{C}_{obs}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = etc$. This would give the same answer.) >>

B. Test the null-hypothesis, $p_1 = p_2$ versus the alternative, $p_1 \neq p_2$ using the (approximate) level of significance, 5%, and interpret the result.

<< **Answer:** The null hypothesis $p_1 = p_2$ is equivalent with the hypothesis $p_1 - p_2 = 0$. Hence, according to supplementary exercise 5 (on the net), we obtain a test with (approximate) level of significance 5% by rejecting H_0 if 0 does not belong to the CI for

$p_1 - p_2$ calculated in section **A**. Since the CI in **A** does not cover 0, we can reject H_0. We can also claim that $p_1 > p_2$ (see supplementary exercise 5) >>

Appendix: Summary of the multinomial distribution

We have *n* independent trials. In each trial one and only one out of *k* categories, $C_1, C_2, \ldots, C_k$, must occur. For every *j*, the probability, p_j, that category C_j occurs, is constant in all trials. Let X_j be the number of times C_j occurs in *n* trials. Then $X_1 + X_2 + \cdots + X_k = n$. The joint distribution of $X_1, X_2, \ldots, X_k$ is called *the multinomial distribution* with parameters, $(n, p_1, p_2, \ldots, p_k)$. The probability mass function (pmf) for $X_1, X_2, \ldots, X_k$ is

$$f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_k = x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where all $p_j \geq 0$ and $p_1 + p_2 + \cdots + p_k = 1$, and where $x_1, x_2, \ldots, x_k$ are integers ≥ 0 such that $x_1 + x_2 + \cdots + x_k = n$. If *k*=2, the distribution reduces to X_1 being binomially distributed (n, p_1) setting $p_2 = 1 - p_1$ and $x_2 = n - x_1$. If $k = 3$, the distribution is often called *the trinomial distribution*.

Some properties:

- The marginal distribution of X_j is binomial (n, p_j) for $j = 1, 2, \ldots, k$, which implies $E(X_j) = np_j$ and $\text{var}(X_j) = np_j(1 - p_j)$.
- The covariance between X_i and X_j is $\text{cov}(X_i, X_j) = -np_i p_j$ for $i \neq j$.
- Unless there are further restrictions on the parameters, the mle for p_j is given by $\hat{p}_j = \frac{X_j}{n}$.